逆文書頻度を利用したグラフベースキーワード抽出手法の提案

逆文書頻度を利用した グラフベースキーワード抽出手法の提案

GRAPH BASED KEYWORD EXTRACTION USING INVERSE DOCUMENT FREQUENCY

上村健流
Takeru Kamimura
指導教員　藤井章博

法政大学大学院理工学研究科応用情報工学専攻修士課程

In recent times, due to the excessiveness of research paper in the web, there is a need of automatic document summarization for papers. And, text summarization is depend on keyword extraction. In this paper, we propose keyword extraction method combining linguistics and graph theory proposed in related work. By inputting the co-occurrence, position, the term frequency, and inverse document frequency of word, this method provide words feature score.

***Key Words** : keyword extraction, text summarization*

１．序論

（１）背景

近年，世界の論文数は増加傾向にあり，2000 年に比べ現在は約 2 倍になっている．また，電子化が進み多くの論文がインターネットを通して読めるようになった．一方，論文数の増加に伴い本来リサーチャーが必要とする論文の取得が困難になっている．それに伴い，関連性のある論文を自動的に収集するシステムの需要が増している[1]．そのためには，大量の文書からテキストを抽出し，適切に処理することで新たなる価値を生み出す必要がある．その手段として，文書の要約，自動的なキーワード抽出などが挙げられる．キーワードの抽出を自動的に行うことができれば，人手を介さずに文書中の特徴的な単語やフレーズを予測することができる[2]．

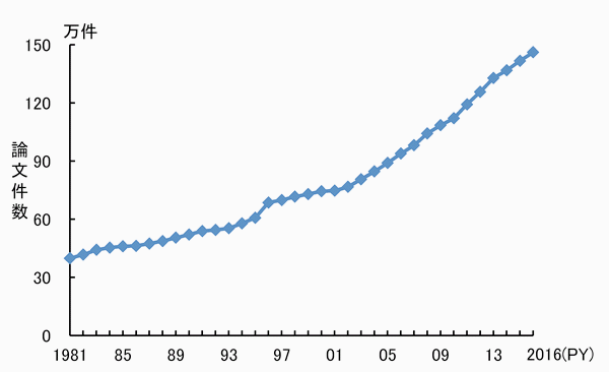

図 1 全世界の論文数の推移[3]

文書の要約は大きく分けて，教師なし手法と教師あり手法の 2 通りに分かれる．教師なし手法では，言語学を基盤として元の文書からテキストを抽出し，読者に提示する．一方教師あり手法では大量の文書を用いて内容を学習しコンピューターによって解釈した表現で提示する．教師あり手法は文書要約では教師なし手法よりも優れたパフォーマンスを発揮するが，文書から直接キーワードを抽出し，タグとして付与したコーパスの必要性が注目されている．

（２）目的

論文中より内容を要約したキーワードを抽出しタグ付けするタスクは，文書分類，情報検索システム，検索エンジンの最適化，曖昧な単語の除去などにおいて重要な役割を果たす．本研究では，教師なしアプローチによるキーワード抽出手法として注目されているグラフベース手法を参考にし，言語学の観点で重要語を特定する手法を組み合わせることで，精度の高いキーワード抽出手法を提案する．また，自然言語処理において英語と日本語にはスペースの有無による形態素解析精度や文法の違いなど，結果に影響する違いがある．そこで本研究では日本語の論文を対象にして評価を行うことで，英語論文に対して有効とされている既存のキーワード抽出手法が日本語に対しても有効であるかどうかの検証も合わせて行う．

２．関連研究

（１）TextRank

2004 年，Empirical Methods in Natural Language Processing にて「TextRank: Bringing Order into Texts」という論文が Rada Mihalcea らによって発表された[4]．Google の検索エンジンが Web ページの重要度を

推定する際に使用する PageRank アルゴリズムをテキストの要約に用いたグラフベースの手法． グラフ理論の中心性の特徴を利用し，キーワードの抽出を行う．PageRank では，Web ページをノード，ページ間のリンクをエッジとしてグラフを生成する．一方 TextRank では単語をノード，単語同士の共起をエッジとしてグラフを生成する．

このランキングアルゴリズムの基盤となる考え方は，ノードからノードへの遷移時の重みの概念である．より重要度の高い候補単語とエッジで結ばれている候補単語(共起している単語)もまた，重要度の高い単語であるという考え方を反映している． また，グラフ内に巡回経路ができてしまい抜け出せなくなる場合を考慮し，一定の確率で他のノードに転移させることで精度が向上する． 以降，ノードから隣接するノードへ移動することを「遷移」，ノードからランダムなノードへ移動することを「転移」と呼ぶ．なお，何単語以内に出現する単語同士を共起とみなすかを示すウィンドウサイズは 6 とする．

G = (*e*, *v*)を頂点集合 *v* とエッジ集合 *e* を持つ有効グラフとする．ここで，頂点 v_iに向かう頂点の集合を $In(v_i)$，頂点 v_iが指す頂点の集合を $Out(v_i)$とする．このとき頂点 v_iのスコア $S(v_i)$は以下のように定義する．

$$S(v_i) = \alpha \sum_{v_l \in In(v_i)} \frac{w_{ij}}{\sum_{v_k \in Out(v_j)} w_{jk}} S(v_j) + (1-\alpha)$$

α は 0 から 1 の値を取る係数であり，ダンピングファクターと呼ばれる．与えられた頂点からグラフ上のランダムに選んだ別の頂点にジャンプする確率を表す. α の値は 0．85 に設定されるのが一般的である[5]．なお，PageRank アルゴリズムにおいてのエッジはリンクを表すため単方向であったが，TextRank アルゴリズムのエッジは単語間の共起を表すため双方向となる．

各ノードの収束がしきい値を下回るか任意の回数を超えるまで，上記の関数を再帰的に呼び出す．以上によって導き出されたノードの値が各単語の重要度を示すスコアとなる．

（２）PositionRank

2017 年， Annual Meeting of the Association for Computational Linguistics(ACL)にて TextRank と同じくグラフ理論に基づいたキーワード抽出手法を提案する「PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents」が Corina Florescu らによって発表された[5]．基本となる考え方は TextRank と同じく，重要な単語の近くに発生した単語もまた重要であるという発想である．さらに，最初に出現する単語は後に出現する単語よりも重要度が高いという仮定を元に，PositionRank では単語の共起確率だけでなく出現頻度，出現場所のパラメータをアルゴリズムに取り込み，精度の評価を行っている． キーワード候補としては名詞，または形容詞と連結した名詞を 1 つのフレーズとして選択する．単語を連結した場合は，各単語のスコアを加算し 1 つのスコアとする．

TextRank ではノードの遷移中に一定確率でランダムなノードに転移するが，PositionRank ではその転移先の選定確率に各単語の位置情報，出現頻度の値を乗算する手法を提案している．単語の出現場所は，出現番号の逆数を用いて算出する． 1 番目の単語は 1/1， 2 番目の単語は 1/2， 3 番目の単語は 1/3 という値を出現した回数分加算していく． 以上の手法によって得られる出現位置，出現頻度によるスコアリングを行う論文要旨の例を図 2 に示す．

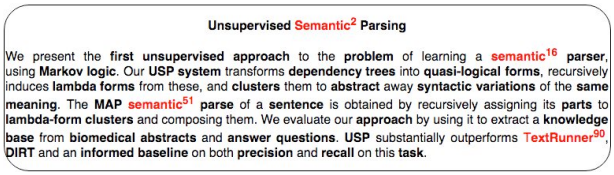

図 2 論文要旨例

図 2 では，例として「semantic」「textrunner」という 2 つの単語の出現位置，出現頻度スコア算出方法に注目する． 「semantic」という単語は 2 番目，16 番目，51 番目に，「textrunner」という単語は 90 番目に出現しているため，p(v_i)を単語 v_i の出現位置，出現頻度スコアとすると，

$$p(semantic) = \frac{1}{2} + \frac{1}{16} + \frac{1}{51} = 0.582$$

$$p(textrunner) = \frac{1}{90} = 0.011$$

となる．

この手法によって得られた文書中の全単語候補の集合を正規化したベクトルを *P* とする．*n*=| *V*| とすると，

$$P = \begin{pmatrix} \frac{p(v_1)}{\sum_{i \in n} p(v_i)} \\ \frac{p(v_2)}{\sum_{i \in n} p(v_i)} \\ \vdots \\ \frac{p(v_n)}{\sum_{i \in n} p(v_i)} \end{pmatrix}$$

最後に，ベクトル *P* を TextRank の転移先ノードの確率に用いて，単語 v_iのランキングスコア $S(v_i)$を以下のように定義する．

$$S(v_i) = \alpha \sum_{v_l \in In(v_i)} \frac{w_{ij}}{\sum_{v_k \in Out(v_j)} w_{jk}} S(v_j) + (1-\alpha) p_i \qquad (1)$$

当該論文では International Conference on Knowledge Discovery and Data Mining(KDD)で発表された論文 834 セット，International World Wide Web Conference(WWW)で発表された論文 1350 セットなどを

対象として評価実験を行っている．なお，各論文は要旨の先頭にタイトルを連結させて使用している．

評価結果は，tf-idf や TextRank などの既存のキーワード抽出手法よりも高い精度が報告されている．また，ウィンドウサイズの値は，精度に影響を与えないことが当該論文で報告されている．

３． 日本語論文における精度評価

日本語の論文に対して PositionRank を適用し，同じく教師なし学習手法として高い精度でキーワードを抽出することができる tf-idf と比較することで性能を評価した．国立情報学研究所が運営する CiNii より，自然言語処理で検索しヒットした論文 1123 件を対象とし名詞，または形容詞と連結した名詞をキーワードの候補とした．PositionRank と tf-idf の日本語論文に対する精度比較結果を図 3 に示す．

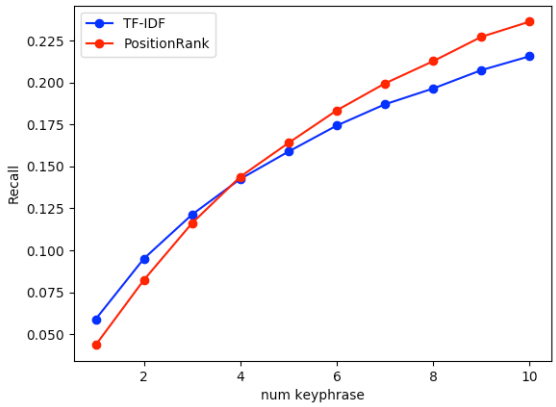

図 3 日本語論文に対する tf-idf と PositionRank の比較

横軸は特徴語上位何単語を取得するかを示す値であり，縦軸は著者が提示したキーワードのうち，本実験で取得できた特徴語と一致しているものの割合を示す再現率である．図 3 より，上位 4 つ以上の特徴語を抽出した際に PositionRank の精度が tf-idf を上回った．

４． 提案手法

特徴語抽出アルゴリズムである PositionRank を参考にし，新たなアルゴリズムを提案する．本研究では，以下の 3 つの手法を試し精度を評価した．「単語の出現位置による重みを緩やかにする手法」「転移先単語の重みに逆文書頻度を反映する手法」「隣接行列に逆文書頻度を反映する手法」の合計 3 つの手法を試し，精度を評価した．

（１）単語の出現位置による重みを緩やかにする手法

PositionRank において単語の出現位置によるバイアスの値は，単語の出現位置の逆数を使用していた．しかし，この手法では序盤に出現する単語間のスコアリングの差と，後半に出現する単語間のスコアリングの差に違いが発生してしまう．例えば，1 単語目と 2 単語目の値の差(1/1 - 1/2 = 0．5)に比べて 50 単語目と 51 単語目の値の差(1/50 - 1/51 = 0．0004)は小さくなってしまう．そこで，単語の出現位置による重みを緩やかにするため，出現位置の値に対数をとり各単語のスコアを算出するアルゴリズムで特徴語の抽出を行った．図 2 の論文要旨における「semanteic」という単語を例に挙げると，以下のような値になる．

$$p(semantic) = \frac{1}{\log 2} + \frac{1}{\log 16} + \frac{1}{\log 51} = 2．058$$

P の値を定めた後，ウィンドウサイズを 6，ダンピングファクターを 0．85 として数式(1)を使用して各単語のスコアリングを行う．

（２）転移先単語の重みに逆文書頻度を反映する手法

PositionRank では 1-α=0.15 の確率で転移する先の重みに単語の出現位置，出現回数のバイアスを加えた．本手法では，この重みに tf-idf にて使用される逆文書頻度の値を乗算することで精度の向上を図った．

まず，tf-idf について述べる．tf(term frequency)とは，ある単語が文書 d 中に現れる回数を，文書 d 中のすべての単語数で割った値を意味する．

$$tf(t,\ d) = \frac{n_{t,\ d}}{\sum_{s \in d} n_{s,\ d}}$$

idf(inverse document frequency)とは，ある単語の文書頻度である df(document frequency)の逆数である．ある単語が出現する文書数である df の逆数に対数を用いて算出するのが一般的である．N は全文書数とする．

$$idf(t) = \log \frac{n}{df(t) + 1}$$

最後に tf と idf の積を算出することで，文書中においてどれだけ重要な単語であるかを表す尺度を導き出すことができる．

PositionRank における重みにはすでに単語の出現頻度のバイアスがかかっているため，本手法では逆文書頻度を表す idf の値を使用する．単語 v_i の idf 値を $idf(v_i)$ とすると，本手法による転移先単語の重みベクトル P は以下のようになる．

$$P = \begin{pmatrix} \frac{p(v_1)idf(v_1)}{\sum_{i \in n} p(v_i)} \\ \frac{p(v_2)idf(v_2)}{\sum_{i \in n} p(v_i)} \\ \vdots \\ \frac{p(v_n)v_n}{\sum_{i \in n} p(v_i)} \end{pmatrix}$$

（３）隣接行列に逆文書頻度を反映する手法

4．2 の手法では，ノードの転移が発生したときのみ idf の値を利用した．そのため，隣接行列 W に沿ってノード間を移動する際には TextRank，PositionRank など従来の手法と同じ経路を辿ることになる．そこで，idf の値を転移先単語の重みではなく，隣接行列 W に乗算し，単語のスコアリングを行った．隣接行列 W は単語間を移動するときに元の単語のスコアに乗算される値であるが，本手法では更に移動元の単語の idf 値を W に乗算する．

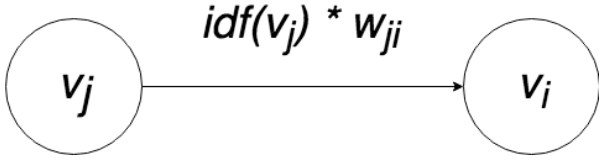

図 4 提案手法の隣接行列モデル

このモデルを適用した場合，単語 v_i のスコア $S(v_i)$は

$$S(v_i) = \alpha \sum_{v_l \in In(v_i)} \frac{idf(v_j) w_{ij}}{\sum_{v_k \in Out(v_j)} w_{jk}} S(v_j) + (1-\alpha) p_i$$

と表すことができる．本手法ではこのアルゴリズムを用いて各単語のスコアリングを行う．

５． 既存手法との比較

（１）提案手法の精度評価

a）単語の出現位置による重みを緩やかにする手法の評価

単語の出現位置による重みを緩やかにする手法の評価結果を図 5 に示す．

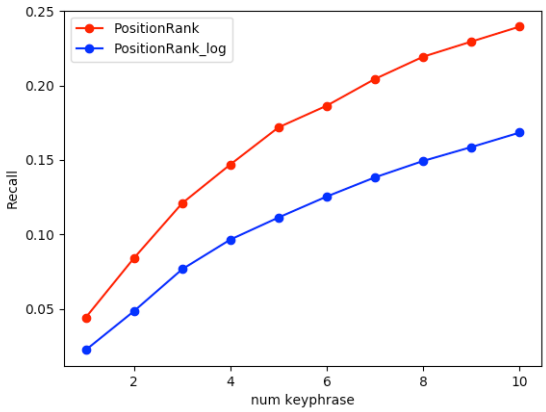

図 5 単語位置バイアスを緩やかにする手法の評価結果

図 5 において，PositionRank_log が本手法の結果を表している．単語出現位置の重みに対数を適用し差を緩やかにした結果，PositionRank の精度を下回った．

b）転移先単語の重みに逆文書頻度を反映する手法の評価

転移先単語の重みに逆文書頻度を反映する手法の評価結果を図 6 に示す．

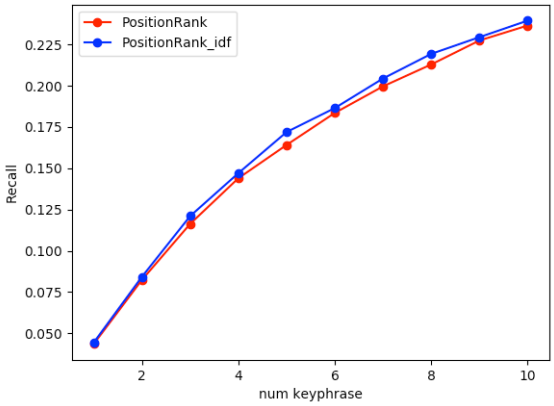

図 6 p に idf を反映する手法の評価

図 6 において，PositionRank_idf が本手法の結果を表している．転移先単語の重みに逆文書頻度を乗算した結果，僅かに PositionRank の精度を上回った．

c）隣接行列に逆文書頻度を反映する手法の評価

隣接行列に逆文書頻度を反映する手法の評価結果を図 7 に示す．なお，転移先単語の重みに逆文書頻度を反映する手法も比較対象とする．

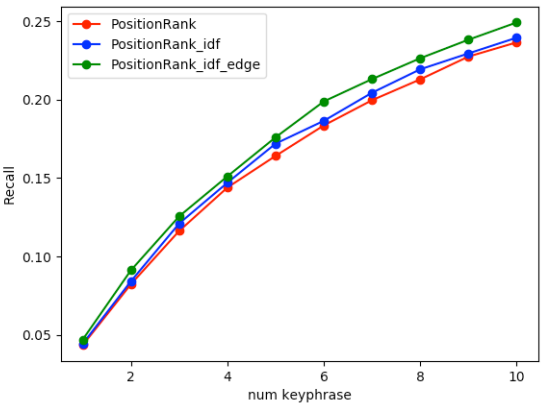

図 7 隣接行列に idf を反映する手法の評価

図 7 において，PositionRank_idf_edge が本手法の結果を表している．隣接行列の各要素に逆文書頻度を乗算する手法は，上記手法や PositionRank の精度を上回った．

再現率の高かった，転移先単語の重みに逆文書頻度を反映する手法，隣接行列に逆文書頻度を反映する手法をそれぞれ PositionRank_idf，PositionRank_idf_edge として既存手法との比較結果を表 1 に示す．

表 1 既存手法との比較

データセット	手法	再現率(%)			
		上位2単語	上位4単語	上位6単語	上位8単語
自然言語処理論文	tf-idf	9.5	14.3	17.4	19.7
	PositionRank	8.2	14.4	18.4	21.3
	PositionRank_idf	8.4	14.7	18.6	21.9
	PositionRank_idf_edge	9.1	15.1	19.9	22.6

６． 統計的有意性

上記の結果だけでは 2 つの提案手法が既存の手法を上回ったという結果が統計的に有意であると断定することはできない．そこで，t 検定を行うことで有意性を調べた．

t 検定とは，2 つの母集団それぞれから抽出した標本の平均に差があるかどうか算出するを検定である．2 つの母集団から平均と分散を元に，どの程度差が生じ得るかを表す t 値を求めることによって，得られたデータの希少性を表す p 値を導くことができる．ここで，母集団 1 の平均値を$\overline{x_1}$，不偏分散を s_1，サンプルサイズを n_1 とすると t 値は以下のように求められる．

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

本項では，自然言語処理論文 1123 件のデータセットを対象として，各手法と既存の手法である PositionRank の精度が等しいという帰無仮説を設定し，t 値ならびに p 値を求めることで有意水準 0．05 を下回るか判定した． なお，t 検定は既存手法の精度を上回った 2 つの手法に対してのみ行った．

（１）転移先単語の重みに逆文書頻度を反映する手法の有意性

転移先単語の重みに逆文書頻度を反映する手法と PositionRank の有意差を表す t 検定結果を図 8 に示す．

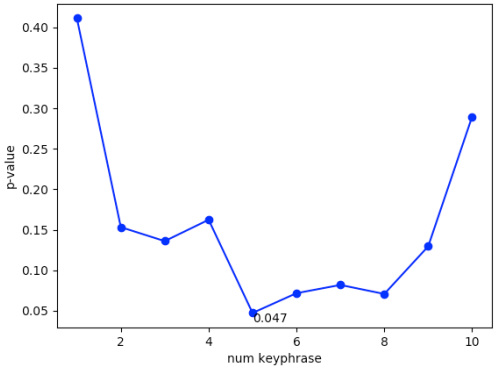

図 8 転移先単語の重みに idf を反映する手法の t 検定結果

図 8 より，num keyphrase が 5 のときに p 値が 0．05 を下回っている．よって，提案手法と PositionRank の間には特徴語上位 5 つを抽出した際の再現率に統計的有意差があると言える．

（２）隣接行列に逆文書頻度を反映する手法の有意性

接行列に逆文書頻度を反映する手法と PositionRank の有意差を表す t 検定結果を図 9 に示す．

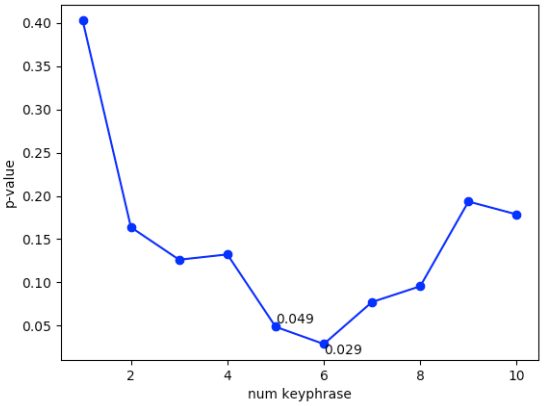

図 9 隣接行列に idf を反映する手法の t 検定結果

図 9 より，num keyphrase が 5，6 のときに p 値が 0．05 を下回っている．よって，提案手法と PositionRank の間には特徴語上位 5 つ，または 6 つを抽出した際の再現率に統計的有意差があると言える．

７． 考察

（１）日本語論文に対する PositionRank の考察

PositionRank の元論文では英語の論文に対して精度評価を行い，tf-idf など既存の手法よりも優れていることを示した．本研究では PositionRank によるキーワード抽出が日本語の論文に対しても有効であるか確かめるために評価実験を行った．その結果，上位 4 つ以上の特徴語を抽出する精度で PositionRank が tf-idf を上回った．一方，上位 1 つ~3 つの特徴語を抽出する精度は tf-idf の方が高かった．その理由としては，英語と日本語の文法的違い，単語間のスペースの有無による形態素解析精度の違いなどが考えられる．また，本研究では「りんご」と「林檎」などの漢字と平仮名による表記揺れに対応していないことも原因の 1 つと考えられる．そのため，表記ゆれをカバーしたコーパスを用意し単語の正規化精度を高めることで，日本語の論文に対するキーワード抽出精度の向上が期待できる．

（２）各提案手法に関する考察

PositionRank を参考に，新たに 3 つの手法を示し評価を行った．

1 つ目の, 単語の出現位置による重みを緩やかにする手法では既存手法の精度を大きく下回った．このことから，序盤に出現する単語の重要度のバイアスを大きく設定している PositionRank のアルゴリズムが適切なものであることが分かった．

2 つ目の, 転移先単語の重みに逆文書頻度を反映する手法では既存手法の精度を僅かにに上回った．このことから，逆文書頻度の値はグラフ理論による特徴語抽出タスクとの共存が可能であることが分かった．

3 つ目の, 隣接行列に逆文書頻度を反映する手法では既存手法や上記の手法よりも高い精度を示した．転移先単語の重みに逆文書頻度を反映する手法では，逆文書頻度の値を乗算するのは一定確率で発生するノードの転移のときのみであった．一方，本手法ではノード間を遷移する場合に参照する隣接行列に逆頻度文書の値を乗算した．グラフ理論における重要な要素である中心性の考え方は，ノード間の遷移を繰り返す部分に反映されているため，本手法では中心性に対してより強く逆文書頻度の影響与えることができたと考えられる．しかし，逆文書頻度の値は母集団によって変化するため，必ずしも今回のような結果が得られとは言えない．そのため，本手法がどんな母集団に対して有効であるかを検証する必要もあると考えられる．

謝辞：本研究を進めるにあたり，ご指導頂いた藤井章博教授に感謝いたします．また，共に切磋琢磨してきた藤井研究室の皆様に感謝いたします．

参考文献

1) S. Nam Kim, O. Medelyan, M. Kan, and T. Baldwin, "Automatic keyphrase extraction from scientific articles" Language Re-sources and Evaluation, Springer 47(3), 2013, pp. 723–742.

2) C. Zhang, "Automatic keyword extraction from documents using conditional random fields" Journal of Computational Information Systems, vol. 4 (3), 2008, pp. 1169-1180.

3) “科学技術・学術政策研究所科学技術指標 2018”, http://www.nistep.go.jp/sti_indicator/2018/RM274_41.html , 2019/2/13 アクセス

4) R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts" Empirical Methods in Natural Language Processing, 2004, pp. 404-411.

5) C . Florescu and C . Caragea , "PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents" Annual Meeting of the Association for Computational Linguistics, 2007, pp. 1105-1115.